DOCUMENT RESUME

ED 161 934 TM 007 961

AUTHOR Berk, Ronald A.
TITLE Item Sampling From Finite Domains of Written Discourse.
PUB DATE Mar 78
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Readability; *Sampling; Sentences; *Test Construction; *Test Items; *Transformation Generative Grammar
IDENTIFIERS Bormuth Literacy Model; *Sample Size

ABSTRACT

A method is described for choosing sample sizes when the domain to be sampled consists of a finite set of sentences and the purpose is to construct a test to assess the comprehension or the readability of written discourse. The testing method is that proposed in Bormuth's work on transformational analysis within a criterion-referenced measurement framework. The suggested sampling unit for this type of analysis is the sentence, but other units are mentioned. Guidelines are provided for computing sample size based on: the size of the domain, the criterion level for successful performance, and the confidence level desired. Two methods are suggested for selecting the sample of sentences to which the transformational rules for item writing are to be applied. A table of sample sizes for a given set of conditions is presented to aid the practitioner in using this sampling procedure. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED161934

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald A. Berk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ITEM SAMPLING FROM FINITE DOMAINS OF WRITTEN DISCOURSE

Ronald A. Berk

The Johns Hopkins University

Paper presented at the annual meeting of the
American Educational Research Association

Toronto, March 1978

Reproduced in U.S.A.

TM007 961

Abstract

A sampling methodology is proposed for determining the lengths of tests designed to assess the comprehension or readability of written discourse. It is an extension of Bormuth's (1970) work on transformational analysis within a criterion-referenced measurement framework. The sampling units specific to this type of analysis and alternative sampling models are examined. Guidelines are provided for computing sample size and selecting the sample of sentences to which the transformational rules can be applied. A table of sample sizes for a given set of conditions is presented to aid the practitioner in utilizing the sampling theory.

Among the recent advances in the measurement of literal comprehension has been the development of a scheme for transforming sentences into test items. It is known as transformational analysis. Derived from theoretical linguistics by Bormuth (1970), transformational analysis represents one of the first major attempts to operationally define domains of written discourse such that the generation of finite item domains is possible. Sets of explicit rules are employed to transform sentences selected from textual material into items that measure comprehension of those sentences. The rules provide a direct link between the content and the items.

The analysis begins with a designated body of content. It must be in a verbal and print format such as a literal transcript of instruction or a textbook. The content is then broken down into its component elements. Initially, this may take the form of passages or paragraphs. Each of these elements must be further reduced to sentences — simple, compound, or complex. The "sentence" is the only unit amenable to item transformations. The number of sentences in a basal reader, for example, might constitute a domain from which items can be generated. Transformational rules are then applied to the domain in order to transform each "item of content" into a test item. They are based on the theory of transformational grammar (Chomsky, 1965) and, therefore, pertain specifically to the semantic and syntactic relationships of the language.

Probably the most familiar transformations are the echo, tag, yes/no, wh-, paraphrase, and semantic substitute. These are only a few of many that can be used to transform a base sentence into one or more items. Descriptions of 60 different transformations have been compiled by Menzel (1970). Bormuth (1970) emphasizes that

> ...populations of sentence-based questions can be completely defined and taxonomized by specifying (a) the syntactic nodes tested, (b) the position of the nodes in the deep structure and surface

structures of sentences, (c) the operations performed on the base sentences, (d) the item transformation applied, and (e) the operations used to select the response distractors when they are used. (p.50)

While transformational analysis provides a methodology for creating a domain of items, in practice, only a sample of these items need to be produced. A random sampling procedure should be followed in selecting the items that comprise the final test. This ensures the sample is unbiased and representative of the domain. Unfortunately, despite the importance of sampling in the test design, the transformational analysts have devoted little attention to the procedure. Other than mentioning that it must be random, they do not offer any specific guidelines for determining how many passages and/or sentences should be selected to adequately assess an individual's comprehension of the domain.

This paper describes an application of the sampling theory commonly used to solve survey research problems to the topic of sentence sampling. It will furnish a theoretical foundation for answering questions about test length in the measurement of comprehension and readability. Given the nature of the content domains and the decisions often made in regard to an individual's competencies in reading, the sampling procedures will be developed according to a criterion-referenced test framework. They are organized into four sections: (1) sampling unit, (2) sampling method, (3) sample size, and (4) sample selection.

Sampling Unit

In transformational analysis, three potential sampling units can be identified. These units and their respective domains are defined below:

Sampling Unit	Domain Size
Passages or paragraphs from a test	Number of passages in the test
Sentences within each passage	Number of sentences in the passage
Nodes within each sentence	Number of nodes in the sentence

Sampling Method

The sampling units suggest at least three alternative sampling plans: (1) one-stage cluster sampling of passages; (2) two-stage or subsampling of passages and sentences within passages; (3) one-stage simple random sampling of sentences. The first method involves the random sampling of passages from the passage domain defined above. All sentences within each of the selected passages would constitute the sentence sample. The second method requires the passage sampling of the first method plus a random sampling of sentences from each sampled passage. The third method combines the passage and sentence domains into one domain consisting of the total number of sentences in the text. One sentence sample is then drawn from that domain.

Simple random sampling has several distinct advantages over the other two approaches. First, it is statistically more efficient. This is based on the characteristics and size of the passage and sentence domains in the textual material commonly available. Secondly, the computations required to arrive at the final sample are simpler and less time-consuming. Finally, the method precludes the task of passage sampling.

Once a sentence sample is selected, the sampling of structure and lexical nodes is relatively easy. Since there are only a small number of nodes in any single sentence, no sample size computations are necessary. One node can be randomly drawn from among the five to 10 nodes that comprise the domain.

Sample Size

There are five major factors that affect sample size: (1) the statistic used as the basis for interpreting the test results; (2) the estimation error associated with the sample statistic; (3) the level of confidence one can place in that statistic; (4) the magnitude of the statistic; (5) the domain size. Each of these will be described in relation to sentence sampling. Since at least one item is typically transformed from each sentence, it will be assumed for the purposes of sample size

computation that sentence sampling is equivalent to item sampling. No mathematical derivations of the formulae will be presented. They can be found in most any standard volume on the topic (e.g., Cochran, 1963; Hansen, Hurwitz, & Madow, 1953; Kish, 1965).

Proportion Correct (Criterion level)

In criterion-referenced measurement the proportion of items a student must answer correctly to attain mastery of a content domain is the statistic of interest. It serves as the basis for both individual and group decision-making. While the criterion level of comprehension performance can be set in a variety of ways, the method used to arrive at that criterion has no bearing on the determination of sample size. What is important is the relationship between the criterion of domain performance and the criterion on the sample of sentences/items drawn from that domain. For example, is 80 percent correct specified as a domain criterion equivalent to 80 percent correct applied to any sample selected from that domain? The answer is only coincidentally. The criterion in each case cannot be presumed to be identical. The criterion applied to a single sample randomly selected from a domain provides an estimate of the domain criterion. There is always some margin of error associated with the sample criterion. The task is to assess and control for this error in the sampling process.

Error Limit

A test-maker can either sample any number of sentences/items from a domain and compute the amount of error post hoc or systematically plan into the sample selection a limit on the amount of error he/she is willing to incur. The former approach is inadvisable in that the resultant error may be unacceptable, thereby rendering the sample unreliable. In addition, the approach does not take advantage of the mechanism for controlling the error prior to the selection which is readily accessible. Therefore, it is recommended that the error limit be prespecified and incorporated into the computation of sample size.

The error limit is determined judgmentally rather than statistically. Two factors must be taken into account in making this judgment: (1) the intended use of the criterion and (2) the costs of achieving various levels of precision. Since the criterion is the point at which most decisions are made, the margin of error about the criterion should be minimized. Ten percent error related to a criterion of 80 percent would mean that the domain criterion will likely fall within a range 10 percentage points lower (70 percent) to 10 percentage points higher (90 percent). This range is obtained by simply adding and subtracting the percentage of error ($\pm$10 percent) to the criterion. If the 20 point interval is felt to be too wide, it can be narrowed by reducing the allowable error.

The precision corresponding to different error limits varies inversely with the sample size. In general, as the error decreases the sample size increases. This is the cost one must pay for greater precision in estimating the criterion. As the domain size becomes large, however, the incremental increases in the sample size become smaller (see Table 1). This demonstrates a basic precept of probability sampling — the principle of diminishing returns with increases in absolute sample size (Warwick & Lininger, 1975). The reductions in sampling error decrease as the absolute size of the sample increases.

Confidence Level

Specified in conjunction with the error limit is the probability or level of confidence one can place in knowing that the domain criterion will fall within the interval mentioned previously. The width of the interval is a function of both the error limit and the probability level. For any given error and criterion, the greater the width, the higher the probability. The most frequently used level is 95 percent. It indicates that the criterion can be expected to lie within the interval formed by the error limit (70 - 90 percent) 95 chances out of 100. The test-maker can be confident that theoretically if 100 sentence samples were selected from a sen-

tence domain, the criterion will fall within the confidence interval in 95 of these samples and outside the interval in only five of the samples. This extablishes a relatively high level of confidence in the accuracy of the results. Although other confidence levels may be employed, the 95 percent level seems most appropriate for estimating the domain criterion.

When the confidence level is incorporated into the computations for sample size, it is expressed in units of the standard normal curve as opposed to a percentage. The units are standard deviation units. Different confidence levels interpreted in terms of these units (t value) are listed below:

Confidence Level	t Value
68	1.00
80	1.28
90	1.64
95	1.96
99	2.58

Standard Error of a Proportion

In order to avoid confusion in setting error limits and confidence levels, the test-maker should be aware of the difference between the terms error limit and standard error. Consistent with the preceding interpretation of the confidence level, the standard error of a proportion applicable in this case expresses the error in standard deviation units. It is, in fact, the standard deviation of a theoretical distribution of sample proportions. At the 95 percent confidence level and a specific criterion level the standard error approximates one half of the error limit. The estimate of standard error for the preceding example is 5 percent. The confidence interval of 70 to 90 percent is created by computing $\pm$1.96 standard errors (5 percent) around the criterion (80 percent). This is almost $\pm$1 error limit (10 percent). The formula for computing the standard error is:

$$\epsilon_p = \sqrt{\frac{(N - n)}{(n - 1)N} PQ} \ .$$

Where:

ϵ_p = estimate of standard error of a proportion/percentage

n = number of sentences/items in the sample

N = number of sentences/items in the domain

P = criterion level of performance (e.g., 80 or 80%)

Q = 1 - P (e.g., .20 or 20%)

It should become apparent from an examination of the formula that"...increases in the absolute size of the sample do more to reduce the standard error than do comparable increases in the proportion that the sample is to the total population" (Warwick & Lininger, 1975, p.93). The absolute size of a sample is much more important that its proportionate size. A 10 percent sample of sentences (100) from a domain (1000) is not twice as reliable as the 5 percent sample (50). The proportionate size is relevant to the sampling procedure only when it is so large that sampling is virtually pointless. This occurs when the sample size is one half or more of the domain size. A further discussion of this issue will be presented at the conclusion of this section.

If a sample of 58 sentences were randomly selected from a domain of 1000 sentences and the criterion level was specified at 80 percent, the computation of the standard error would proceed as follows:

$$\epsilon_p = \sqrt{\frac{(1000 - 58)\ (.80)(.20)}{(58 - 1)(1000)}} = \sqrt{\frac{942\ (.16)}{57000}} = \sqrt{.0026} = .05$$

The .05 or 5 percent error is interpretable in terms of the domain criterion only by applying a particular confidence level and computing the corresponding confidence interval. This step is necessary when an error limit is not prespecified in the computation of sample size. Otherwise, it may be omitted.

Magnitude of Criterion

Once the error limit and confidence level have been set, the criterion itself needs to be considered. Similar to the error limit, the magnitude of the criterion varies inversely with the sample size — the higher the criterion, the smaller the required sample size. Unlike the error limit, the sample size diminishes symmetrically with increases in the criterion. Criterion levels of performance typically range from 70 to 100 percent. In many instances where the criterion is not known at this stage of development, a conservative level such as 70 percent should be chosen tentatively in order to compute the sample size. This ensures that the sample size will be adequate for that level or higher.

Domain Size

The final ingredient essential for determining sample size is the domain size. This is simply the count of sentences in the domain previously defined. As noted earlier, the sample size increases as the domain size increases. The principle of diminishing returns characterizes the sizes of the sample increases relative to the absolute sample sizes. This trend can be observed in Table 1.

Computation

All of the foregoing elements can now be substituted into the following equation to compute the sample size:

$$n = \frac{\frac{t^2PQ}{e^2}}{1 + \frac{1}{N}\left(\frac{t^2PQ}{e^2} - 1\right)}$$

Adapted from Cochran's formula 4.1 (1963, p.75)

Where:

n = number of sentences/items in the sample

t = 95% confidence level expressed as a t value (1.96)

P = criterion level of performance (e.g., .80 or 80%)

Q = 1 - P (e.g., .20 or 20%)

e = error limit (e.g., .10 or 10%)

N = number of sentences/items in the domain

The values of P, Q, and e can be expressed as proportions or percentages. When percentages are used, the quantity Q is equal to 100 - P. An application of the formula based on percentages is shown below:

Given:

$t = 1.96$

$P = 80$

$Q = 20$

$e = 10$

$N = 1000$ sentences

$$n = \frac{\frac{(1.96)^2(80)(20)}{(10)^2}}{1 + \frac{1}{1000}\left(\frac{(1.96)^2(80)(20)}{(10)^2} - 1\right)} = \frac{61.4656}{1.0605} = 58$$

An inspection of the formula and computations reveals how the sample size is adjusted according to the values of the five elements. It increases directly with t and N and inversely with P, Q, and e.

To aid the practitioner in determining sample size for a given set of conditions, Table 1 has been prepared. Sample sizes are computed for seven common criterion levels ranging from 70 to 99 percent. The 99 percent level should serve as an approximation to the 100 percent level. It is mathematically impossible to compute sample sizes for that criterion. The error limits, confidence level, and domain sizes were selected in terms of their appropriateness for sentence/item sampling.

Insert Table 1 about here

The obvious disparity between the sample sizes required for a 5 percent error limit and those required for a 10 percent error limit should illustrate the price one must pay for greater precision. With regard to the former, in particular, the proportionate size of the sample relative to the domain size is quite large when the

domain contains fewer than 500 sentences and the criterion level is low. The utility of sampling should be seriously questioned whenever the computed sample size is more than one half the size of the domain. Under those circumstances the transformational analysis may have to be conducted on the entire domain. The practical constraints governing a particular sampling problem will likely dictate the exact sample size to be drawn.

Sample Selection

There are a number of procedures one can use to draw a random sample from a domain. The most popular involves the selection of a set of random numbers equal to the sample size. This may be done manually or by computer. If a sample of 58 sentences were to be selected from a domain of 1000 sentences, for example, 58 random numbers between 0001 and 1000 would be chosen, i.e., 559, 385, 306, 69, ... The 58 sentences corresponding to these numbers would comprise the sample. Tables of random numbers are available in most statistics books.

A far simpler and more convenient approach equivalent to the above is recommended for sentence sampling. It is called systematic sampling. The equivalence of the two procedures rests on the assumption that the sentence domain can be considered random with respect to the variable being measured — the sentence structure. The justification for this assumption follows.

In a sentence domain such as a basal reader, the sentences are not arranged in any order that would suggest that the structure of any sentence in the first story may be distinctly different from any sentence appearing in a later story. While new vocabulary is often introduced with each successive story, there is no reason to suspect the more difficult words will markedly affect the sentence structures. Consequently, one can assume the sentences to be essentially random in terms of their structures.

The first step in the systematic sampling of sentences is to determine the interval k, so that every k^{th} sentence is selected from the domain. It is obtained by dividing the domain size (N) by the computed sample size (n), or

$$k = \frac{N}{n} .$$

For the sample size of 58 and the domain size of 1000, k equals 17.

Starting with a randomly selected sentence among the first 17, every 17th sentence in the domain would then be identified. Suppose the first number between 1 and 17 chosen from a table of random numbers was 8. The sample would consist of the sentences numbered 8, 25, 42, 59, ..., 943, 960, 977, 994.

The use of the sampling interval ensures in a systematic fashion that the sample will not only exhaust the entire domain but will be representative of it. Proceeding by intervals is certainly much easier than assigning random numbers, particularly when the sample size and/or domain size is large.

Other variations of the approach when a recurring pattern of sentence structures in the domain is evident are described by Cochran (1963). Their relationships to stratified and cluster sampling are explained by Madow (1946) and Madow and Madow (1944).

Conclusions

The proposed sampling methodology provides a theoretical basis for determining the lengths of tests designed to assess the comprehension or readability of written discourse. Practically, it extends the work on tranformational analysis by offering guidelines for sampling sentences/items from the domains of textual material commonly available. This would facilitate the application of transformational rules to produce tests that measure comprehension of those materials.

Despite the utility of the sampling procedures, several technical issues remain unresolved. For instance, how many and what type of transformations should be performed on a sentence that is sampled from a given domain have not been delineated.

Other concerns related to the operations of transformational analysis have also been expressed from the perspectives of psycholinguistics (Perfetti, 1972) and psychometrics (Cronbach, 1970; Diederich, 1970; Mehrens, 1970). Certainly a considerable amount of attention needs to be devoted to these problems in order to fully operationalize item transformations and realize their potential for measuring literal comprehension.

Table 1

Sample Sizes Required for Estimating Criterion Levels of 70 to 99% in Selected Domains at the 95% Confidence Level (Upper and lower numbers designate the minimum sample sizes for error limits of 10% and 5%, respectively).

	Criterion Level						
Domain Size	70	75	80	85	90	95	99
50	31	30	28	25	21	14	4
	43	43	42	40	37	30	12
100	45	42	38	33	26	16	4
	77	74	71	66	58	42	13
250	61	56	50	41	31	17	4
	141	134	124	110	89	57	14
500	70	63	55	45	32	18	4
	197	183	165	141	109	64	15
750	73	66	57	46	33	18	4
	226	208	185	156	117	66	15
1000	75	67	(58)	47	34	18	4
	244	224	198	164	122	68	15
5000	79	71	61	49	34	18	4
	304	273	234	189	135	72	15
10000	80	72	61	49	35	18	4
	313	280	240	192	136	73	15
50000-Infinity	80	72	62	49	35	18	4
	323	288	246	196	138	73	15

Example: The number of sentences that need to be randomly sampled from a domain of 1000 sentences with a criterion level of 80% and an error limit of ±10% at the 95% confidence level is 58. This means that 95 times out of 100 the domain criterion can be expected to lie within the interval 70-90%.

REFERENCES

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: M.I.T. Press, 1965.

Cochran, W. G. Sampling techniques (2nd ed.). New York: Wiley, 1963.

Cronbach, L. J. Review of Bormuth's On the theory of achievement test items. Psychometrika, 1970, 35, 509-511.

Diederich, P B. Review of Bormuth's On the theory of achievement test items. Educational and Psychological Measurement, 1970, 30, 1003-1005.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. Sample survey methods and theory: Volume I, methods and applications. New York: Wiley, 1953.

Kish, L. Survey sampling. New York: Wiley, 1965.

Madow, L. H. Systematic sampling and its relation to other sampling designs. Journal of the American Statistical Association, 1946, 41, 207-214.

Madow, W. G., & Madow, L. H. On the theory of systematic sampling. Annals of Mathematical Statistics, 1944, 15, 1-24.

Mehrens, W. A. Scientific test construction: Pure and sterile (Review of Bormuth's On the theory of achievement test items). Contemporary Psychology, 1970, 15, 666-667.

Menzel, P. The linguistic bases of the theory of writing items for instruction stated in natural language. In J. R. Bormuth, On the theory of achievement test items. Chicago: University of Chicago Press, 1970. Pp. 87-155.

Perfetti, C. A. Psychosemantics: Some cognitive aspects of structured meaning. Psychological Bulletin, 1972, 78, 241-259.

Warwick, D. P., & Lininger, C. A. The sample survey: Theory and practice. New York: McGraw-Hill, 1975.